Evaluation of Feature Channels for Correlation-Filter-Based Visual Object Tracking in Infrared Spectrum

Erhan Gundogdu[1], Aykut Koc[1], Berkan Solmaz[1], Riad I. Hammoud[2], A. Aydın Alatan[3,4]
[1]Intelligent Data Analytics Research Program Dept., Aselsan Research Center, Ankara, Turkey
[2]BAE Systems, Burlington, MA, USA
[3]Center for Image Analysis (OGAM), [4]Electrical and Electronics Eng. Dept., METU Ankara, Turkey
[1]{egundogdu,aykutkoc,bsolmaz}@aselsan.com.tr, [2]hammoud@mit.edu, [3]alatan@eee.metu.edu.tr

Abstract

Correlation filters for visual object tracking in visible imagery has been well-studied. Most of the correlation-filter-based methods use either raw image intensities or feature maps of gradient orientations or color channels. However, well-known features designed for visible spectrum may not be ideal for infrared object tracking, since infrared and visible spectra have dissimilar characteristics in general. We assess the performance of two state-of-the-art correlation-filter-based object tracking methods on Linköping Thermal InfraRed (LTIR) dataset of medium wave and longwave infrared videos, using deep convolutional neural networks (CNN) features as well as other traditional hand-crafted descriptors. The deep CNN features are trained on an infrared dataset consisting of 16K objects for a supervised classification task. The highest performance in terms of the overlap metric is achieved when these deep CNN features are utilized in a correlation-filter-based tracker.

1. Introduction

Visual object tracking is an active area in computer vision research, and most tracking methods have been widely applied to visible spectrum imagery to serve as a useful tool of various tasks including human computer interface [34], action recognition [29, 16] and robotics [36]. State-of-the-art methods such as support vector machines [17], sparse dictionary-based learning [3, 30, 27], saliency-based approaches [40] and correlation filters [5, 10, 18] have been applied in an online fashion for object tracking. Moreover, extensive benchmarks have been presented such as Online Tracking Benchmark (OTB) 2013 [39], Visual Object Tracking (VOT) 2014 [23] and VOT 2015 [22] challenges with different evaluation metrics. Experimenting on these diverse datasets, recent methods mostly focus on the tracking routine itself rather than the features/descriptors used for tracking. Tracking methods adopt the latest features of object classification, most of which are based on the gradient orientations, and tuned considering the visible spectrum characteristics. A recent study [12] has already compared the newest methods in an infrared benchmark dataset [4] with conventional features, which might not be feasible for night time imagery applications. To this end, the need to analyze feature types is vital in infrared spectrum. Hence, in this work, we provide a fair comparison on Linköping Thermal InfraRed (LTIR) dataset [4], including 20 sequences collected from various sources and annotated in the VOT-toolkit [23, 22] format.

Similar to object classification, the discriminative and representative power of the features play a crucial role in object tracking. Since online tracking applications require both a robust performance and an efficient implementation, correlation filters, being computationally efficient and reliable, have become the prevailing approach for visual tracking. The seminal work of Bolme et. al. [5] introduced a fast implementation, where they used raw image intensities. Following their work, multichannel correlation-filter-based tracking methods have been proposed using histogram of oriented gradients (HOG) [7] features [18, 10], and color channels [11]. Although HOG feature channels improve the performance significantly in visible imagery compared to the raw intensities, we show that these features, in conjunction with specific tracking methods, perform weakly in infrared images. Moreover, unlike color images, infrared images capture information on a single band (e.g. medium wave infrared (MWIR), longwave infrared (LWIR)) unless a multispectral sensor is utilized. In the study of [15], the performances of Haar-like features and image intensity-based features are gauged in a visible/infrared comparison context. Their major conclusion is that Haar-like features cause a dramatic performance degradation when the spectrum is changed from visible to infrared, compared to using

Approved for public release; unlimited distribution.

the raw image intensities.

The issues mentioned above motivate us to propose the use of alternative feature types. Although there exist plenty of features that could be integrated into a correlation-filter-based tracking approach, we compare three feature types, which we believe are the most beneficial ones for object tracking in infrared imagery. The features that we compare involve (1) those hand-crafted features such as Gist-feature maps (where each map represents the band-pass response of the visual object) and gradient orientation maps (i.e. HOG feature channels) (2) the infrared feature maps, which are extracted using the first-layer weights of CNN filters learned independently for classification on a separate infrared dataset, and which we call as "deep infrared signatures (deep-IRS)". To evaluate the role of these feature types on visual object tracking task in infrared imagery, we adopt correlation-filter-based tracking methods DSST [10] and SRDCF [9] due to their recent success on Visual Object Tracking Challenges [23, 22, 12].

In this work, our contributions are as follows: (1) we evaluated the top-performing correlation trackers of both VOT 2014 [23] and VOT 2015 [22] challenge, (2) two robust features are integrated into these two recent methods and (3) one of the compared feature-map types is extracted using infrared-specific filters learned by a CNN architecture for an infrared classification task. (4) Extensive analyses are carried out using the VOT-toolkit [22, 23] on LTIR dataset [4].

In the rest of the paper, we first present the mostly related works of object tracking divided into four broad categories since complementary approaches to correlation-filter-based approaches are also worth to discuss. Next, we provide a brief explanation of the adopted correlation-filter-based algorithms in Sec. 3. In Sec. 4, evaluated feature types are presented, and the experimental results and discussions are provided in Sec. 5.

2. Related Work

2.1. Generative approaches

Generative models construct an appearance model, which is updated when necessary. Tracking is accomplished by searching for the most-likely object candidate within a search area on the next frame. Incremental visual tracking (IVT) [33], mean shift tracking [6] and visual tracking decomposition [24] are some prominent examples of robust and generative tracking methods. In the context of generative models, sparse representations have also been used in visual tracking [30, 3, 27], since such representation improves the face recognition [38], due to its robustness against occlusion and noise corruption.

2.2. Discriminative approaches

In discriminative approaches, tracking is achieved upon a learned binary classifier (e.g. target object vs. background) and the most probable target location is estimated using classification. The pioneering studies are tracking with online boosting [14] and ensemble tracking [1], where both use an online version of AdaBoost [13] and update weak classifiers according to the object location. Multiple Instance Learning-based tracking (MILTrack) [2], Online Discriminative Feature Selection (ODFS) [43], Fast Compressive Tracking (FCT) [44] improve the performance of feature-based discriminative tracking methods by the help of Haar-like features. In [17], an online structured-output support vector machine (SVM) is designed for visual tracking.

2.3. Correlation-filter-based methods

Instead of using a discriminative classifier, correlation-filter-based trackers [5, 19, 18, 25, 10] employ dense correlation in the image domain to localize the object. Minimum Output Sum of Squared Errors (MOSSE) [5] learns a filter which minimizes an objective function aiming at obtaining a sharp peak in the correlation mask. [19] exploits the circulant structure of the cyclic shifts of a signal and applies kernel regression. The main drawback of the correlation-filter-based trackers [5, 19] is that they are not scale-adaptive. Hence, the tracking performance often degrades in video sequences when the size of the target is subject to large variations. To handle this situation, multiscale solutions are proposed such as Discriminative Scale Space Tracker (DSST) [10], Scale Adaptive Multiple Features (SAMF) tracker [25], spatio-temporal context (STC) tracker [42]. Since the standard correlation-filter-based methods suffer from the periodic boundary effect (which leads to imperfect training samples and a restricted search region), Spatially Regularized Discriminatively learned Correlation Filters (SRDCF) [9] is proposed to alleviate these sufferings.

2.4. Hybrid Methods

As opposed to using a single approach, hybrid methods employ complementary approaches and combine them to compensate for their individual drawbacks. For instance, multiple correlation trackers are run at different parts of the object in [28] whereas part-based and holistic methods are combined in [37]. Reliable patches are tracked in [26] using Kernelized Correlation Filters (KCF) [18] as the base tracker. The work in Multiple Experts using Entropy Minimization (MEEM) [21] selects the best support vector machine-based discriminative tracker according to an entropy minimization criterion. Like MEEM, an ensemble-based method is proposed in [15], which selects the best

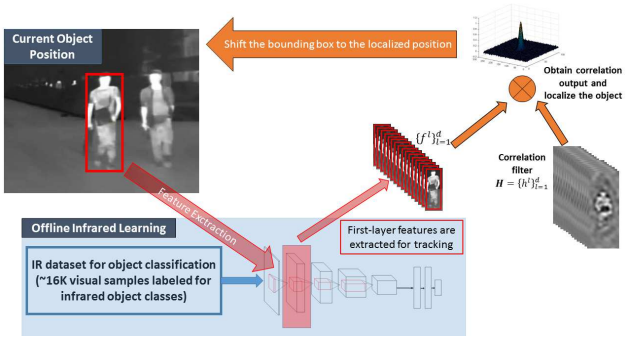

Figure 1. Illustration of the correlation-filter-based tracking overview with deep-IRS features.

expert from the ensemble using a sparse appearance dictionary, and a corresponding template exists for each tracker in the ensemble. In [20], Markov Chain Monte Carlo sampling selects trackers and combines them. Various trackers with mixed feature types are combined in [35]. Hybrid methods combining generative and discriminative approaches are proposed in [45, 41].

3. Multichannel Correlation-Filter-Based Visual Tracking

Most correlation-filter-based tracking methods compute the correlation of the candidate object patch f around the current region of interest with a template filter h, and the location which gives the highest correlation score within a search rectangle (i.e. gate) is determined as the estimated location of the object at the current frame [5, 10, 18, 19]. The fast correlation is achieved in the frequency domain using the Convolution Theorem as $G = F \odot H$, where lower and uppercase letters denote the signal in image and frequency domain, respectively, and $\odot$ denotoes the element-wise product. By taking the inverse Discrete Fourier Transform of G, the correlation output is obtained. Since Fast Fourier Transform (FFT) is used for finding the corresponding signals in the frequency domain, the correlation calculation has a complexity of $O(Plog(P))$ rather than $O(P^2)$ (P denotes the discrete signal length). Unlike using only raw image intensities of the template f, most state-of-the-art methods employ more robust features [18, 19, 10] extracted from the template f (such as HOG orientation maps or any other set of feature maps), and these extracted features are $\{f_i^1, ..., f_i^d\}$ processed to be correlated in the same way as described above. An overview of the correlation-filter-based tracking algorithm is presented in Figure 1 for the case of deep infrared signature features (deep-IRS).

In this work, the top performing correlation-filter-based tracking methods of VOT 2014 [23] and VOT 2015 [22] are used as the baseline tracking approaches. The overall winner of VOT 2014 is DSST [10] whereas the top performing correlation method of VOT 2015 is SRDCF [9]. The following subsections briefly summarize these methods.

3.1. Discriminative Scale Space Tracker (DSST)

DSST [10] is an extension of MOSSE [5] where multichannels (HOG channels) are utilized for localization in addition to the raw image intensities. The desired correlation mask of the training example f is denoted by g and the ridge regression cost in Eq (1) is intended to be minimized:

$$\sum_{i=1}^{t} \left\| (\sum_{l=1}^{d} h_t^l * f_i^l - g_i) \right\|^2 + \lambda_\epsilon \sum_{l=1}^{d} \left\| h_t^l \right\|^2 \qquad (1)$$

Here, $*$ denotes the circular correlation operation, λ_ϵ is the control parameter for L_2 regularization term of the filters $\{h_t^l\}_{l=1}^d$, and $\{f_i^1, ..., f_i^d\}$ are the feature channels which may correspond to particular features (HOG, Gist, IR features). There exists a closed form solution to the minimization of the regularized cost in (1) for $t = 1$, i.e., for one training example, in frequency domain as:

$$H^l = \frac{F^l \odot G_*}{\sum_{k=1}^{d} F^k \odot F_*^k + \lambda_\epsilon}, \forall l \in 1, ..., d, \qquad (2)$$

where $\odot$ is the element-wise product and the subscript $_*$ denotes conjugation operation. At each time instant, the filter H^l is updated by applying moving average to the numerator and denominator of (2) separately via:

$$\begin{aligned} A_t^l &= (1-\mu) A_{t-1}^l + \mu G_{t*} \odot F_t^l, \\ B_t &= (1-\mu) B_{t-1} + \mu \sum_{k=1}^{d} F_t^k \odot F_{t*}^k, \end{aligned} \qquad (3)$$

where μ is the model update rate. The correlation of an object patch z and the model H^l is calculated using the updated numerator A_t^l and denominator B_t^l of H^l in frequency domain and the spatial domain correlation mask is computed by taking the inverse Fourier transform as:

$$y = \mathcal{F}^{-1} \left\{ (\sum_{l=1}^{d} A_*^l \odot Z^l) / (B + \lambda_\epsilon) \right\}, \qquad (4)$$

The new location of the object in the next frame is found using (4). Scale estimation is performed on the translated location. Features are extracted in the translated location using variable patch sizes. The same correlation filtering procedure is employed in the scale space using these features (cf. [10] for details) to find the accurate scale of the object in the next frame.

3.2. Spatially Regularized Discriminatively Learned Correlation Filters

Since correlation filters suffer from the limited search range and imperfect training examples due to the cyclic shifts, SRDCF [9] adds an extra term w to their cost function in Eq. (1):

$$\sum_{i=1}^{t} \left\| (\sum_{l=1}^{d} h_t^l * f_i^l - g_i) \right\|^2 + \lambda_\epsilon \sum_{l=1}^{d} \left\| w \odot h_t^l \right\|^2, \qquad (5)$$

where w penalizes the values of the correlation filter h_t^l's at the image boundaries by using a quadratic function that has low values near the center of the image patch and high values near the boundaries. To solve the optimization problem in (5), they use an efficient optimization procedure. Similar to DSST, they also use HOG feature channels. Yet, the computed feature maps are 16 times smaller than their original area for efficiency since a 4×4 HOG cell size is selected. In our evaluations, we first calculate the feature maps of Gist and deep-IRS with the original patch size and then downscale them by a factor of 4 for a fair comparison with the baseline SRDCF implementation, which uses HOG features.

3.3. Ensemble of MOSSE [5] Trackers

In [15], a comparison study is conducted for visual object tracking on pairs of synchronized visible and infrared sequences. The main comparison is done on the exploited feature type of a tracker (i.e. Haar-like or raw image intensities). The results of the work in [15] claim that the performance of Haar-like features fades away when the spectrum is changed to the infrared. Moreover, they use an ensemble method, named TBOOST, that gives promising results despite exploiting only raw image intensity values. Hence, we also implemented this method and compared it with other methods. The proposed tracker in [15] basically generates a limited dictionary consisting of object templates. Each object template belongs to a different MOSSE [5] tracker. At each frame, only one tracker is run and its output is assigned as the resulting bounding box of that frame. The final tracker to be run is generated by combining all MOSSE filters in the ensemble with different weights. These weights are the coefficients of the sparse reconstruction of the current object patch using the template dictionary. In our comparisons, TBOOST represents the method utilizing only the raw image intensities, since its performance is reported to be significantly higher than that of a single MOSSE tracker.

4. Correlation Feature Channels For Object Tracking

In this work, we propose to evaluate hand-crafted and learned features. As the hand-crafted features, Gist and HOG feature maps are exploited. As learned features, we adopt infrared-specific features (deep-IRS). In addition to these features, experiments with only raw image intensities are also conducted. Figure 2 illustrates the examples from three feature maps: deep-IRS, Gist, and HOG.

4.1. Histogram of Oriented Gradients

HOG [7] is an image representation frequently used in various computer vision tasks, including object detection and classification, since it possesses invariance to a tolerable amount of discrepancies within the same object class. This feature type is also exploited in the correlation-filter-based methods such as [10, 18, 9]. Since HOG involves a global representation (histogram) and lacks the spatial relationships within the object, it is inconvenient to use histograms in a correlation-filter-based tracking framework. To use HOG representations in correlation, the cell size (please refer to [7] for further details about important parameters) is kept so small (typical cell size is 1 pixel) that the resulting elements of HOG bins at each cell represent the gradient orientations for each pixel. Thus, HOG channels exploited in most correlation-filter-based trackers [10, 18, 9] correspond to the gradient orientations maps (with contrast sensitivity and insensitivity). During the experiments, the default parameters of DSST [10] and SRDCF [9] are inherited for the HOG channels.

4.2. Deep Infrared Signatures (Deep-IRS)

An increasing amount of attention to CNN architectures has started to lead researchers to use CNN features in different research problems. Transferring the features trained in one domain to another domain is also a commonly used strategy, especially when the domains are close to each other [8]. Features learned by training a CNN architecture has already been employed in the correlation-filter-based tracking [8]. In [8], the most useful CNN layer is experimentally determined to be the first-layer feature maps since the translational invariance properties and spatial information are preserved. Following this work, we utilize IR-specific feature maps.

To learn these IR features and to increase the performance of classical classification and detection methods in infrared spectrum, we generate a dataset consisting of IR objects captured by MWIR (3-5um) and LWIR (8-12um) cameras. Sample object patches are demonstrated in Figure 3. For the classification task, the dataset consists of 16K labeled object samples for different object categories including pedestrian, ship, vehicle, airplane and helicopter. The

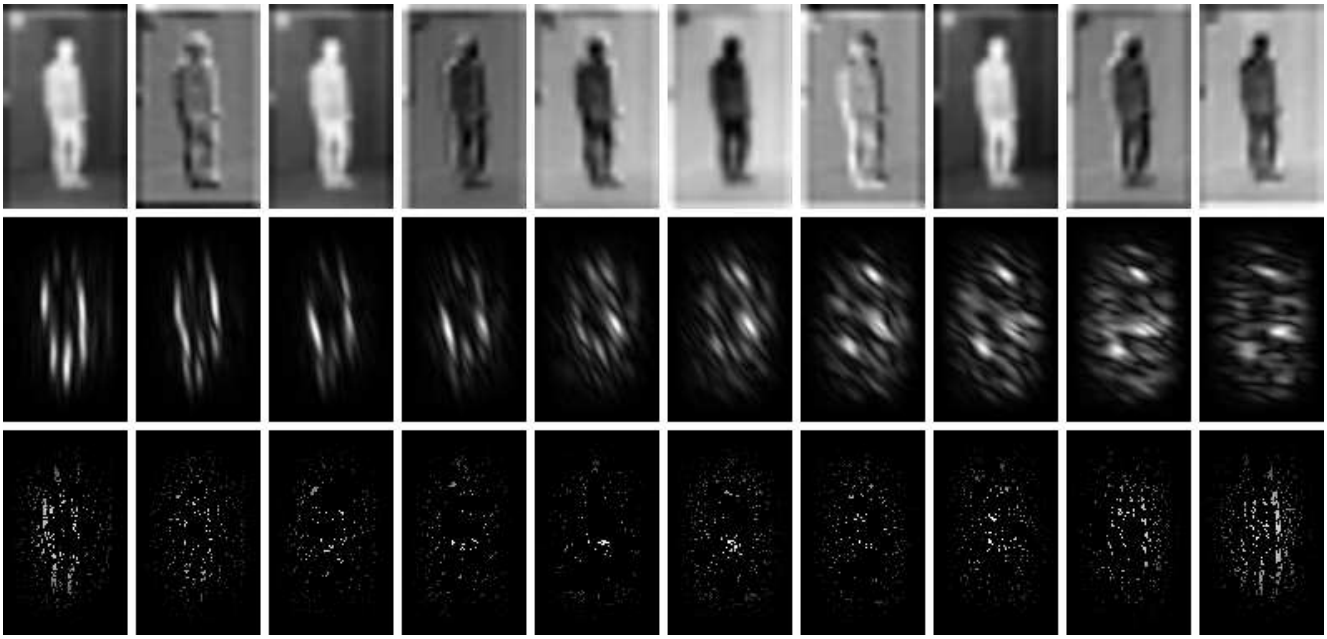

Figure 2. A visual illustration of features on the *Hiding* sequence of LTIR [4] dataset: **Top row**: infrared feature maps extracted by the prelearned deep CNN first layers (deep-IRS). **Middle row**: Gist feature maps, where each one corresponds to a response to the band pass filters. **Bottom row**: HOG feature maps with each one representing a gradient in a specific orientation. Intensity values of the feature channels are normalized between 0 and 1 for better illustration.

employed CNN architecture is as follows: $(64 \times 64) \rightarrow (5 \times 5 \times 20) \rightarrow (5 \times 5 \times 40) \rightarrow (4 \times 4 \times 80) \rightarrow (5 \times 5 \times 512) \rightarrow (512 \times 512) \rightarrow (512 \times 5)$ using conventional neural network notation. For each convolutional block, there is a rectified linear unit (ReLU), a dropout (with a factor of 0.5) and a 2×2 max pooling layer with a stride 2. To extract the IR feature maps, we utilize the first-layer responses of the CNN architecture that is trained on the generated IR dataset. During the training part, no images from the LTIR dataset [4] are used. Hence, during the training and testing, the images, which are captured in distinct environments by the help of the sensors with different properties, are processed. This makes our comparisons and evaluation realistic for a practical usage.

4.3. Gist features

The Gist, proposed by Oliva et al. [32, 31], is a holistic scene descriptor based on power spectrum features, and is well known for its effectiveness in scene classification. In our experiments, we utilize power-spectrum-based maps for enhancing the tracking performance in infrared imagery. The images containing the tracked objects are prefiltered with the aim of reducing illumination effects (which is also encountered in thermal imagery as dynamics change and temperature change [4]) such as large shadows in the scene as well as preventing local and high-contrast image regions

Figure 3. Sample image patches from our infrared dataset.

that disturbs the image power spectrum. First, a logarithmic function is applied to the intensity distribution, very low spatial frequencies are attenuated by the use of a high-pass filter. Next, the local standard deviation at each pixel of the image is adjusted to make large regions of the image appear equally bright. These operations are given in Eq. 6,

$$i'(x,y) = \frac{i(x,y) * h(x,y)}{\epsilon + \sqrt{[i(x,y) * h(x,y)]^2 * g(x,y)}} \qquad (6)$$

where $i(x,y)$, $g(x,y)$ and $h(x,y) = 1 - g(x,y)$ represent the image, an isotropic low-pass gaussian spatial filter, and the high-pass filter that removes the mean intensity value of the image and whitens the energy spectrum, respectively. The denominator is simply a local estimator of the variance

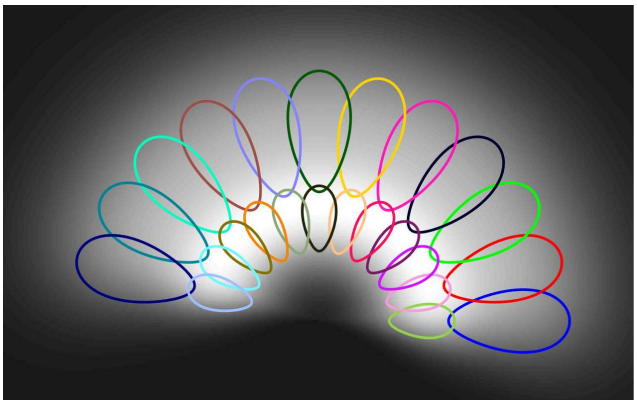

Figure 4. Visualization of filters. Each colored blob is a Gabor filter in the frequency domain. The outer set of 12 blobs represents the first scale, and the inner set of 12 blobs represents the second scale.

of the output of the high-pass filter and is used for noise reduction.

After prefiltering, power spectrum is computed for each image by applying a Discrete Fourier Transform (DFT), and is sampled with a bank of narrow-band Gabor filters. The transfer function of each filter, tuned to a spatial frequency, f_r, is:

$$G(f_x, f_y) = Ke^{-2\pi^2(\sigma_x^2(f_x^{'}-f_r)^2+\sigma_y^2 f'^2_y)} \quad (7)$$

where $f_x^{'} = f_x cos(\theta) + f_y sin(\theta)$, $f_y^{'} = -f_x sin(\theta) + f_y cos(\theta)$ define the filter orientation, and σ_x and σ_y define the filter shape. The filter bank is a set of filters with different orientations and scales as depicted in Fig. 4. For Gist feature channels, we generate 2 scales of filters for DSST, each scale consists of 12 filters. Since SRDCF downsamples the object patches by a factor for 4, we selected a more hierarchical configuration, 3 scales of filters, each consisting of 16, 8 and 7 filters, respectively. These filters span all frequency spectrum uniformly.

5. Experiments and Results

5.1. LTIR dataset

We perform experiments on Linköping Thermal InfraRed (LTIR) dataset [4], captured with cameras of MWIR and LWIR bands, for evaluating the tracking methods combined with different feature types. LTIR consists of 20 sequences including various targets such as rhinoceros, horses, humans, dogs, quadrocopters and cars. Since their annotations are provided in VOT-toolkit format, we evaluate the methods in terms of VOT metrics, which are explained in Section 5.2. The dataset includes both indoor and outdoor scenarios with different attributes such as dynamics change, temperature change, blur and camera motion. Average length of a sequence is 563 frames.

5.2. Evaluation metrics

In our evaluation, the performance metrics of VOT 2014 [23] and 2015 [22] are utilized. These are the average accuracy, robustness scores, the ranking measure of [23] for VOT 2014 dataset and expected overlap measure, which is proposed in VOT 2015 challenge [22].

For a predicted object region and its ground truth at frame t, accuracy is defined as $Acc_t = \frac{area(R_P \cap R_T)}{area(R_P \cup R_T)}$, where R_P and R_T represent the predicted and the true object regions in a specific frame, respectively. Average accuracy per sequence is calculated by averaging these accuracy scores through time. If a tracker fails, i.e., accuracy score decreases to zero, then the tracker is re-initialized (please refer to VOT 2014 challenge paper [23] for further details). On the other hand, robustness measures the number of failures per frame. Ranking of a tracker for each performance metric is calculated by ordering each tracker among the evaluation set, merging the tracker orders sharing statistically very similar results, and finally averaging these rankings for all the sequences.

Since the accuracy and robustness are complementary measures, a new metric, named *expected overlap*, is proposed in [22] that takes the average of accuracy scores in a principled manner to unify robustness and accuracy metrics. This measure has been used as the new metric in VOT 2015 (please refer to [22] for details). To calculate the expected average overlap ratio of a tracker, several tracking segments are stored according to their length. For each tracking segment length, an expectation is calculated and the average of all of the expectations from different segments constitutes the final performance score.

5.3. Experimental results and discussion

Table 1 summarizes the performance results in terms of the accuracy ranking, the robustness ranking, the average accuracy, the average robustness and the expected overlap. When we analyze DSST and the usage of different feature types, a performance increase is observed in terms of expected overlap if the feature type is deep-IRS. Moreover, the robustness also increases with the use of deep-IRS features. In addition to these quantitative results in Table 1, we also obtain important visual observations, where DSST with deep-IRS features handle occlusion and large pose variations whereas HOG features fail. Some of the observations from the sequences of LTIR [4] are presented in the supplementary material.

Although SRDCF is the top correlation-filter-based approach of VOT 2015, its performance is lower than DSST in thermal datasets in terms of expected overlap since each image is resized to 16 times smaller that its original area,

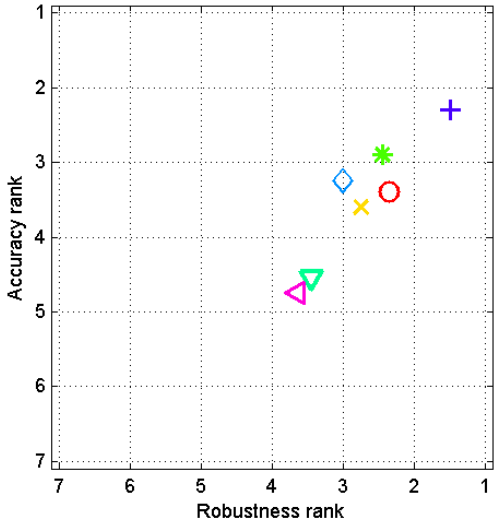

Figure 5. Accuracy and robustness ranking plot. Please refer to Figure 7 for the legend.

and the representation power of the features degrades in the infrared spectrum. Furthermore, neither Gist nor deep-IRS features improves the performance of SRDCF method.

In VOTTIR 2015 challenge [12], the best performing method was SRDCF with an additional feature obtained by subtraction of consecutive frames. Since our goal is to compare the algorithms and different feature types, and the frame subtraction dominates the results, we opt to continue without frame subtraction. Moreover, it is not always feasible to use frame subtraction especially when the camera or object has a large motion, though this is not the case in LTIR. Consequently, we avoid overfitting on the dataset and exploiting any prior information while probing the effects of feature types.

Figure 5 illustrates the accuracy and robustness rankings. Since the closeness to the top right region means better performance, SRDCF using HOG features is the best performing one among the compared methods. However, expected overlap measure (shown in Figure 6) considers all different sequence lengths and takes their average. DSST, with deep-IRS features, prevails in terms of this metric, followed by the Gist features. The superiority of deep-IRS features over the HOG orientations in infrared imagery is possibly due to the fact that deep features are learned using an infrared dataset with similar characteristics to LTIR [4] since both datasets are on the same bands (medium wave and long-wave) though the two tasks (tracking and classification in infrared) are dissimilar. Moreover, discriminating properties of HOG channels do not arise in infrared spectrum.

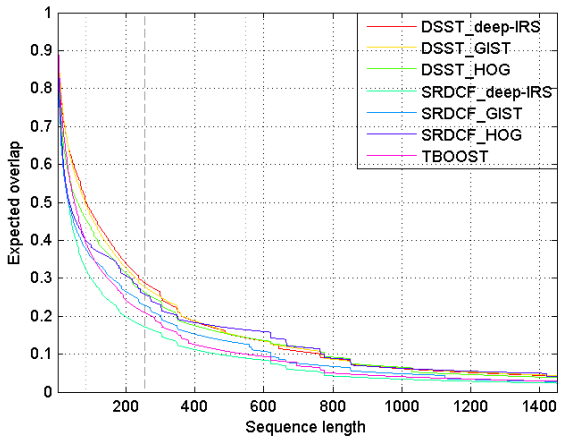

Figure 6. Expected overlap curves.

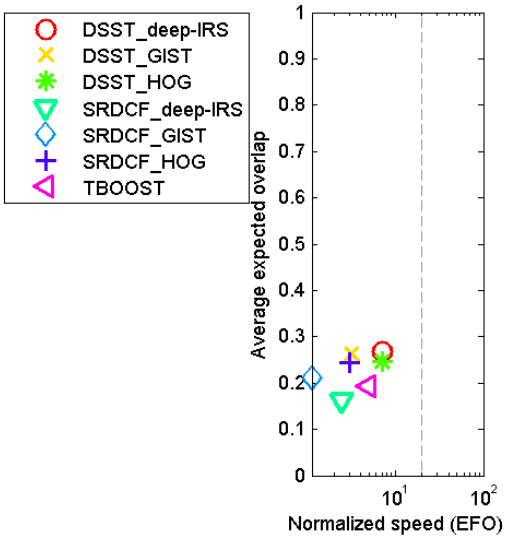

Figure 7. Expected overlap vs normalized speed graph.

Although the accuracy and robustness are important performance metrics, the computation time is crucial from the practical aspects. In Figure 7, the normalized speed is measured in terms of equivalent filter operations (EFO) ([23]) that the processor can run instead of processing the necessary operations for one frame of the corresponding tracker. The most efficient method is DSST with deep-IRS features since the number of features is 20 whereas 28 features maps are used for the HOG version. Moreover, SRDCF and its different versions are significantly slower than DSST family. Finally, the speed of TBOOST ranks between DSST and SRDCF family.

Table 1. Performance results of the evaluated tracking methods. Red, blue and green indicate the best, second and third ranking.

Methods	Acc. Rank.	Rob. Rank.	Accuracy	Robustness	Expected overlap	Normalized speed (EFO)
DSST_deep-IRS	3.40	2.35	0.58	2.30	0.2668	7.12
DSST_GIST	3.60	2.75	0.58	2.50	0.2618	3.16
DSST_HOG	2.90	2.45	0.62	2.35	0.2463	6.98
SRDCF_HOG	2.30	1.50	0.67	1.90	0.2450	2.99
SRDCF_GIST	3.25	3.00	0.60	2.85	0.2115	1.11
TBOOST	4.75	3.65	0.56	3.30	0.1922	4.87
SRDCF_IR	4.55	3.45	0.51	3.65	0.1614	2.45

6. Conclusion

In this work, we investigate the performances of correlation-filter-based trackers when different features are exploited in the infrared domain. To compare the feature types, two state-of-the-art methods, which had the top rankings in VOT 2014 and 2015 challenges, are evaluated. The compared feature types include hand-crafted features (Gist, HOG) and learned features (deep-IRS), extracted by the weights of a deep CNN architecture trained for classifying objects, such as pedestrian, ship, vehicle, airplane and helicopter, on an infrared dataset. DSST with learned infrared features performs favorably against the other feature types in terms of expected overlap, while being the most computationally efficient one among the compared methods.

References

[1] S. Avidan. Ensemble tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):261–271, Feb. 2007.

[2] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 983–990, June 2009.

[3] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust l1 tracker using accelerated proximal gradient approach. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1830–1837, June 2012.

[4] A. Berg, J. Ahlberg, and M. Felsberg. A thermal object tracking benchmark. In *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, pages 1–6, Aug 2015.

[5] D. Bolme, J. Beveridge, B. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2544–2550, June 2010.

[6] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(5):564–575, May 2003.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005.

[8] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. Convolutional features for correlation filter based visual tracking. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[9] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[10] M. Danelljan, G. Hger, F. Shahbaz Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[11] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. van de Weijer. Adaptive color attributes for real-time visual tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[12] M. Felsberg, A. Berg, G. Hager, J. Ahlberg, M. Kristan, J. Matas, A. Leonardis, L. Cehovin, G. Fernandez, T. Vojir, G. Nebehay, and R. Pflugfelder. The thermal infrared visual object tracking vot-tir2015 challenge results. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[13] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, Aug. 1997.

[14] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *Proc. BMVC*, pages 6.1–6.10, 2006. doi:10.5244/C.20.6.

[15] E. Gundogdu, H. Ozkan, H. Seckin Demir, H. Ergezer, E. Akagunduz, and S. Kubilay Pakin. Comparison of infrared and visible imagery for object tracking: Toward trackers with superior ir performance. In *CVPRW*, June 2015.

[16] R. I. Hammoud, C. S. Sahin, E. P. Blasch, B. J. Rhodes, and T. Wang. Automatic association of chats and video tracks for activity learning and recognition in aerial video surveillance. *Sensors*, 14(10):19843, 2014.

[17] S. Hare, A. Saffari, and P. Torr. Struck: Structured output tracking with kernels. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 263–270, Nov 2011.

[18] J. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(3):583–596, March 2015.

[19] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *proceedings of the European Conference on Computer Vision*, 2012.

[20] S. M. Jianming Zhang and S. Sclaroff. Tracking by sampling trackers. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1195–1202, Nov 2011.

[21] F. Jurie and M. Dhome. Meem: Robust tracking via multiple experts using entropy minimization. In *in IEEE International Conference of Computer Vsion*, 2014.

[22] M. Kristan, J. Matas, and et. al. The visual object tracking vot2015 challenge results, Dec ICCV Workshops 2015.

[23] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Čehovin, G. Nebehay, T. Vojir, and G. et al. Fernandez. The visual object tracking vot2014 challenge results, ECCV Workshops, 2014.

[24] J. Kwon and K. M. Lee. Visual tracking decomposition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1269–1276, June 2010.

[25] Y. Li and J. Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II*, pages 254–265, 2014.

[26] Y. Li, J. Zhu, and S. C. Hoi. Reliable patch trackers: Robust visual tracking by exploiting reliable patches. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, June 2015.

[27] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong, and C. Kulikowski. Robust and fast collaborative tracking with two stage sparse optimization. In *Computer Vision ECCV 2010*, volume 6314 of *Lecture Notes in Computer Science*, pages 624–637. Springer Berlin Heidelberg, 2010.

[28] T. Liu, G. Wang, and Q. Yang. Real-time part-based visual tracking via adaptive correlation filters. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[29] W.-L. Lu and J. Little. Simultaneous tracking and action recognition using the pca-hog descriptor. In *Computer and Robot Vision, 2006. The 3rd Canadian Conference on*, pages 6–6, June 2006.

[30] X. Mei and H. Ling. Robust visual tracking using l1 minimization. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1436–1443, Sept 2009.

[31] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, May 2001.

[32] A. Oliva, A. B. Torralba, A. Guérin-Dugué, and J. Hérault. Global semantic classification of scenes using power spectrum templates. In *Proceedings of the 1999 International Conference on Challenge of Image Retrieval*, IM'99, pages 9–9, Swinton, UK, UK, 1999. British Computer Society.

[33] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *Int. J. Comput. Vision*, 77(1-3):125–141, May 2008.

[34] G. Shin and J. Chun. Vision-based multimodal human computer interface based on parallel tracking of eye and hand motion. In *Convergence Information Technology, 2007. International Conference on*, pages 2443–2448, Nov 2007.

[35] F. Tang, S. Brennan, Q. Zhao, and H. Tao. Co-tracking using semi-supervised support vector machines. In *ICCV*, pages 1–8, 2007.

[36] C.-Y. Tsai, K.-T. Song, X. Dutoit, H. Van Brussel, and M. Nuttin. Robust mobile robot visual tracking control system using self-tuning kalman filter. In *Computational Intelligence in Robotics and Automation, 2007. CIRA 2007. International Symposium on*, pages 161–166, June 2007.

[37] L. Čehovin, A. Leonardis, and M. Kristan. Visual tracking using anchor templates. In *Proc. IEEE Winter Applications of Computer Vision Conference (WACV)*, March 2016.

[38] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, Feb 2009.

[39] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *CVPR*, pages 2411–2418. IEEE, 2013.

[40] M. Yousefhussien, N. Browning, and C. Kanan. Online tracking using saliency. In *Proc. IEEE Winter Applications of Computer Vision Conference (WACV)*, March 2016.

[41] Q. Yu, T. B. Dinh, and G. G. Medioni. Online tracking and reacquisition using co-trained generative and discriminative trackers. In *in IEEE International Conference of Computer Vsion*, 2008.

[42] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M. Yang. Fast visual tracking via dense spatio-temporal context learning. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 127–141, 2014.

[43] K. Zhang, L. Zhang, and M.-H. Yang. Real-time object tracking via online discriminative feature selection. *Image Processing, IEEE Transactions on*, 22(12):4664–4677, Dec 2013.

[44] K. Zhang, L. Zhang, and M.-H. Yang. Fast compressive tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(10):2002–2015, Oct 2014.

[45] L. H. Zhong Wei and Y. Ming-Hsuan. Robust object tracking via sparsity-based collaborative model. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 1838–1845, Washington, DC, USA, 2012. IEEE Computer Society.